Published in: Witt, Andreas/Metzing, Dieter (eds.): Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology. - Dordrecht: Springer, 2010. pp. 97-123. (Text, Speech and Language Technology 41)

Chapter 6
Discourse Relations and Document Structure

Harald Lüngen, Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Csilla Puskás

Abstract This chapter addresses the requirements and linguistic foundations of automatic relational discourse analysis of complex text types such as scientific journal articles. It is argued that besides lexical and grammatical discourse markers, which have traditionally been employed in discourse parsing, cues derived from the logical and generical document structure and the thematic structure of a text must be taken into account. An approach to modelling such types of linguistic information in terms of XML-based multi-layer annotations and to a text-technological representation of additional knowledge sources is presented. By means of quantitative and qualitative corpus analyses, cues and constraints for automatic discourse analysis can be derived. Furthermore, the proposed representations are used as the input sources for discourse parsing. A short overview of the projected parsing architecture is given.

Keywords Discourse parsing · Discourse relations · Document structure · Text technology · Linguistic annotations · XML

6.1 Introduction

In the past, several approaches to automatic discourse analysis have been developed as applications of relational discourse theories which describe the semantics of discourse. These approaches are often based on the analysis of discourse connectives as well as morphological and syntactic features. Such surface-oriented strategies are adequate and have yielded good results when applied to the analysis of simple text types like newspaper articles, which are characterised by a limited size and a relatively simple document and syntactic structure. When dealing with more complex text types, however, an analysis of lexis and grammar is not sufficient. Sources of knowledge about discourse and document semantics have to be considered as well.

H. Lüngen (✉)
Justus-Liebig-Universität Gießen, Gießen, Germany
e-mail: luengen@uni-giessen.de

This chapter deals with the linguistic foundations of discourse analysis for a complex text type by the example of scientific journal articles. Its focus is on the contribution of logical document structure, generic document structure and thematic structure to discourse parsing. The modelling and representation of linguistic structures and knowledge sources based on text-technological (XML-based) formalisms and methods is addressed. The representations are used in investigating correlations and interactions between different types of linguistic information and serve as an input to a discourse parsing system.

In the project *SemDok*, which is part of the Research Group *Text-technological modelling of information* funded by the German Research Foundation DFG and scheduled to run in its second phase for three years 2005–2008, a discourse parser for the complex text type "scientific research article" is being developed. Scientific articles exhibit a highly complex document structure (both logical document structures and relational discourse structures are deeply nested) and a relatively large average size in terms of word count. The discourse parser is envisaged in a specific application scenario: It shall be part of an explorative reading system which supports novice students in learning to adopt adequate strategies for reading scientific articles. The system shall have two dimensions: Firstly, it shall provide a tool to support selective and explorative reading and, secondly, it shall function as a learning environment where students can acquire knowledge about the genre "scientific article", its generic text type structure (with categories such as *introduction, method, results* and *discussion*) and possible argumentative strategies and thematic structures. Support for explorative and selective reading shall be based on two mechanisms: highlighting text structures and providing automatically generated link lists to different structural nodes as navigation elements. Highlighting and linking both serve as starting points for the exploration of an article. By offering link lists or by directing attention to highlighted passages, readers are guided to thematically or rhetorically significant parts of a text. Additionally, access to the different structural levels of the text is simplified, as the building plan of the text is made explicit.

Highlighting and linking requires the preprocessing of articles. They must be automatically analysed and annotated on the levels of document structure, text type structure, rhetorical and thematic structure. The automatisation of analysis and annotation is necessary to enable users of the system to upload articles that they themselves consider relevant. The discourse parser developed in the *SemDok* project will automatically add discourse structure annotations and thus allow students a personalised use of the system.

The present chapter is structured as follows: Section 6.2 gives a theoretical overview of the different linguistic levels relevant for the analysis of the relational discourse structure of a scientific article: logical document structure, thematic structure (referential structure, lexical cohesion), and generic document structure. Furthermore, our notion of relational discourse structure, which refers to Rhetorical Structure Theory (RST, Mann and Thompson 1988), is introduced. In Sections 6.3.1

and 6.3.2, the corpus and the layers of annotations that we employ in developing and evaluating the parser, are characterised. Section 6.3.3 addresses additional resources such as the discourse marker lexicon and the inventory of rhetorical relations and describes their representation in XML. The chapter is concluded by a short overview of the architecture of the projected discourse parser and an outlook on future work.

6.2 Linguistic Foundations

6.2.1 Document Structure

The research described in this chapter is based on the assumption that documents can be regarded as complex signs. As complex signs they are built up from smaller units in which these units themselves and their connections are constituted by linguistic and visual mechanisms. These units of a document are complex and elementary segments. Elementary segments are usually rectangular areas, which can be delimited clearly according to certain features and are not put together from segments (e.g. paragraphs or headings). Complex segments are adjacent combinations of segments to which a common document function can be assigned (e.g. sections).

Documents can be regarded as signs with respect to their syntagmatic, their semantic and their pragmatic dimensions. In a syntagmatic perspective, documents can be described by grammars which define the way in which segments can be combined to yield valid documents of a certain type. In a semantic perspective, the meaning of a document is a function of the meanings of its parts and its document type. The combination of elementary segments to form complex segments follows compositional principles. These, however, are activated by the document type assumptions and expectations, which complete the compositionally formed document meaning.

Constitutive units of documents are 2D objects, segments. Segments can almost always be geometrically described as rectangles which cover parts of the document area. Segments are e.g. text blocks, tables, headings, address fields, but also graphics and illustrations, i.e. flat objects which have a recognisable coherent structure and can be described not by linguistic means alone. Tables and lists contain on the one hand linguistically definable structures (e.g. lists can be interpreted as coordination), however, on the other hand, they are specified by geometric and graphic properties at the same time.

Only text blocks, which do not show any further geometric properties apart from the line break, represent purely linguistic objects. Text blocks form the transition between the one-dimensionality of the language and the two-dimensionality of the document by being split up mechanically into lines which fill the segment from top to bottom.

Segments are aggregated in the document area in which semantic connections between the segments are encoded by topology and graphic design. The document

area is restricted, though; it is defined by the restrictions of the medium (size of the printable paper, screen or window etc.). If this does not suffice, document parts are formed so that they can be read in a temporal order one after the other (successive pages of a book, window content which can be scrolled, or window content which is replaced by the activation of a link). In this respect, many documents also have a temporal dimension besides the two spatial ones so that one can talk about documents as of a 2.5-dimensionality.

The syntagmatic structure of text segments has been examined quite extensively in text linguistics. Dependencies between the sentences are established by means of cohesion of different types. The linguistic properties of the syntagmatic level of text segments can be described by rules which permit the sentence syntax to continue above the sentence boundaries. The syntagmatic structure of segments with graphic elements, such as tables, is given by the iconic properties of lines, columns and boxes. These relations can also be described by rules that are based primarily on the cognitive processes of perception. Complex segments and whole documents are formed by the aggregation of segments. Typical is the aggregation of several text segments (paragraphs) to form a text body that is provided with a heading to yield a section. The formation of a complex segment is defined by the adjacent aggregation of the segments in the text area. These syntagmatic properties of documents can be described by rules which resemble those for the formation of sentences; they can be collected in a document grammar. In document grammars, the media-specific conditions of a document are omitted systematically. The necessary page breaks are included in a document grammar no more than line breaks are in the descriptions of segments in a text grammar.

Grammatical dependencies indicate semantic relations. The syntactic structure of a sentence licenses the construction or representation of its meaning in a suitable formalism. There are different approaches to text semantics which presuppose the availability of meaning representations for the individual sentences as well as cohesive means for the representation of the meaning of segments. An example of this is the logical text representation in terms of (S)DRT (Asher and Lascarides 2003). The meaning of a document arises from the composite meanings of the segments contained in it in connection with predefined meaning structures which are activated by document type and text type. To combine the meaning of segments it has to be decided which semantic relations are encoded by a certain configuration of segments (e.g. the semantic relationship between heading and text body). By the document type, a text type is activated which specifies a semantic structure which is valid for all instances of this type, regardless of the meanings specified by the segments. So it is clear from the start, e.g. for a scientific article, that the state of the art, methodological questions or results are represented in certain sections of the document.

Based on speech act theory, different expansions have been suggested on the textual level. Motsch and Viehweger (1991) describe the construction of complex illocutions in texts, Schröder (2003) examines the action structure of texts with the same aim. Following this line of research, document functions can be described in a similar way as complex illocutions.

6.2.2 Relational Discourse Structure

Current text-type independent linguistic discourse theories such as the Unified Linguistic Discourse Model (ULDM, Polanyi et al. 2004a, b), Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003, Asher and Vieu 2005), and Rhetorical Structure Theory (RST, Mann and Thompson 1988, Marcu 2000) describe *discourse structures* as a system of discourse coherence relations that hold between adjacent discourse constituents (spans). Discourse constituents can be either elementary discourse segments or complex discourse segments, the latter are relationally structured themselves. It seems to be generally acknowledged that discourse is structured hierarchically, but it is controversial whether the basic information structure for discourse representation should be a tree or a graph. While SDRT employs graph structures, in ULDM and RST, discourse *trees* with labelled nodes and edges are constructed. Recently, Wolf and Gibson (2005) have put forward linguistic arguments for a graph representation of discourse structures.

In the present project, we adopt the view that a discourse representation is basically a tree structure, which may be enhanced to include re-entrant edges in certain well-defined cases (cf. Lüngen et al. 2006a).

It is also generally accepted that there are two main structural types of discourse relations under which all other relations can be subsumed, namely *subordinating* vs. *coordinating* relations. In RST, these types are called *mononuclear* (or sometimes *hypotactic*) and *multinuclear* (*paratactic*) relations. In a mononuclear relation, one of the elements (text spans) involved has the status of being the *nucleus*, the "more salient, essential piece of information" (Carlson et al. 2001) of the relation. The other ones are labelled the *satellites*, which contain "supporting or background information" (Carlson et al. 2001). Like many authors (e.g. Corston-Oliver 1998, Marcu 2000, Egg and Redeker 2005), we restrict the representation of mononuclear relations to binary trees, i.e. with exactly one nucleus and one satellite. In multinuclear relations, all elements (possibly more than two) are labelled as nuclei.

While in ULDM subordinating and coordinating relations are the only types of relations, the original RST is actually a theory about the nature and diversity of mono- and multinuclear discourse relations, thus a set of 26 so-called *rhetorical* relations and their definitions are introduced in Mann and Thompson (1988).

The fact that all rhetorical relations are either mononuclear or multinuclear and that some (such as EVALUATION and INTERPRETATION) are rhetorically similar, and furthermore that some relations are special cases of other relations (e.g. NON-VOLITIONAL-CAUSE and CAUSE), can be accounted for by grouping relations into *classes* and constructing *taxonomies* over these classes. This has previously been done e.g. by Hovy and Maier (1995) and Carlson and Marcu (2001); see also Goecke et al. (2005). On the one hand, Mann and Thompson (1988) have provided a relation set which is supposed to be text type- and application-independent, on the other hand they stress that the set is open to extension. In practice, depending on a text type and application (e.g. discourse analysis vs. generation), specific subsets or extended sets of relations have been chosen (cf. Hovy and Maier 1995). Many of the RST rhetorical relation types examined in the literature, such as EVIDENCE

or INTERPRETATION, are immediately relevant for our text type, which was one factor that led us to opt for RST-based text parsing. Based on relation sets previously described in the literature as well as on corpus investigations, we have defined an extended relation taxonomy for the *SemDok* project, see Section 6.3.3.2.

Discourse theories also differ in their strategies of *discourse interpretation*, that is, the question of how discourse analysis and the construction of a formal representation of a specific discourse is achieved. In a theory like SDRT, a full-fledged semantic representation of discourse segments is required to perform discourse analysis. Its output then is a logical form, too. In the original conception of RST, text spans comprise plain text, not logical forms. Relational analysis as designed in Mann and Thompson (1988), however, also presupposes knowledge about the meaning of discourse segments as well as goals and beliefs of authors and readers about these meanings. Since a complete and robust automatic semantic analysis of input segments seems not feasible, computational analysis of discourse has often relied on linguistic properties that are more easily obtainable, such as *discourse connectives* and syntactic and morphological features derived from (deep or shallow) grammatical analysis, see the projects described in Corston-Oliver (1998), Marcu (2000), Reitter (2003b), Polanyi et al. (2004a), and cf. also the argumentation in Egg and Redeker (2005). This is also the path that is taken in the *SemDok* project. But since we are dealing with a complex text type, we are also investigating cues for the more global (or macro) discourse structure such as thematic structure and lexical cohesion (lexical chains and anaphoric structure, see Section 6.2.3), logical document structure, and text type structure (Section 6.2.4).

In the extract from our corpus in Listing 6.1,[1] the adverbial discourse connective *z.B.* introduces a mononuclear ELABORATION-EXAMPLE relation where the segment that contains the connective is the satellite. This relation defines a complex discourse segment which is related to the previous segment, which contains the discourse marking conjunction *und*, introducing a multinuclear LIST-COORDINATION relation. The corresponding RST tree is shown in Fig. 6.1.[2] An equivalent discourse dependency tree representation according to Danlos (2005), which better

```
<cds type="block" docIdref="i1161">
  <sds id="s260">
    <eds id="e465">In der Schrift hat die Sprachpflege einen etwas besseren Erfolg
        als im Gespräch gehabt.
    </eds>
  </sds>
  <sds id="s261">
    <eds id="e466">In öffentlichen Dokumenten ist man <dm id="i322" lexid="c6"
        lemma="z.b." pos="ADV">z.B.</dm> darauf bedacht, dass die Termini
        dem Gebrauch in Schweden entsprechen,
    </eds>
    <eds id="e468"><dm id="i325" lexid="c89" lemma="und" pos="CC">und</dm>
        man vermeidet auch typisch finnlandschwedische Wendungen.
    </eds>
  </sds>
  ...
</cds>
```

Listing 6.1 Discourse segments and discourse markers

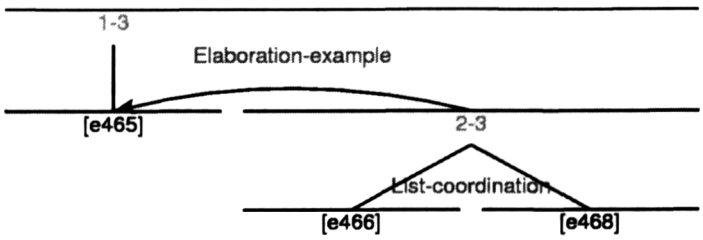

Fig. 6.1 RST tree

corresponds to data structures preferred in computational linguistics, is shown in Fig. 6.2. The involved segments are represented by IDs that refer to the textual content of elementary discourse segments as shown in Listing 6.1.

Elementary discourse segments (EDSs) in our project are based on syntax (syntactic tagging), punctuation and logical document structure. The basic idea is that elementary discourse segments correspond to clauses as in most theories, but may also correspond to other kinds of phrases () when they are especially marked by punctuation (e.g. bracketing) or logical document structure (e.g. a `<doc:title>` element). Moreover, a minimal unit of discourse is supposed to be part of a discourse relation where the nucleus is semantically independent enough so that the satellite can potentially be omitted. This means that e.g. complement clauses, conditional clauses, and restricting relative clauses cannot be EDSs in our scheme. Since in these respects we deviate from the definition of English *elementary discourse units* in Marcu (1999), we did not adopt his technical term *edu* for our minimal segments.

We developed a discourse segmenter that is able to perform EDS segmentation automatically based on the input of the syntactic and logical document structure annotations (annotation layers CNX and DOC, cf. Section 6.3.2) of an input text. It outputs a new annotation layer called SEG, where besides EDSs, also SDSs (sentential discourse segments, i.e. sentences) from the text, and CDSs (complex discourse segments, which correspond to DOC elements) are marked, as can be seen

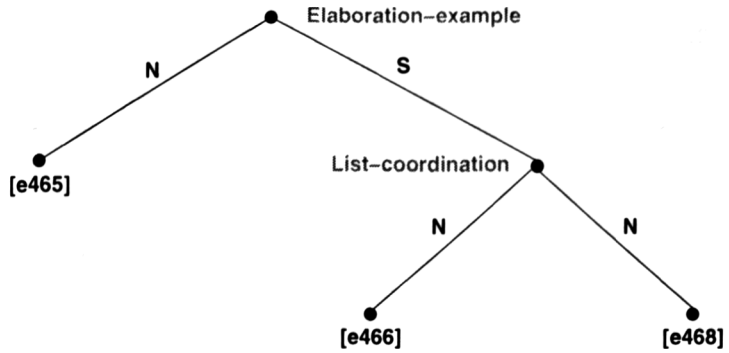

Fig. 6.2 Discourse dependency tree according to Danlos (2005)

in Listing 6.1. The criteria for EDSs as well as the discourse segmenter algorithm are described in Lüngen et al. (2006b).

Among CDSs, we further distinguish three types (cf. Bärenfänger et al. 2006): First, CDS type="block" corresponds to paragraphs and 2D objects that are on a par with paragraphs, such as titles, captions, and images, i.e. the elementary element types from the document structure that contain only text or non-textual 2D objects like images or diagrams, cf. Section 6.2.1. Second, CDS type="division" corresponds to the lowest section level or elements that are on a par with it in terms of DOC markup, e.g. titles and paragraphs that are sisters of section elements. Finally, CDS type="document" comprises all residual section elements, i.e. those which are on a higher level than CDS type="division". In our approach to discourse parsing, these segment types serve to constrain the extent to which discourse segment can be relationally combined, e.g. a CDS type="block" can only be related to another CDS type="block", but not a CDS type="division". In practice this means that the core parser module is called several times in a cascade architecture, starting out with EDSs, and 1each time using the next higher one of the above sketched segments types as its base segment type.

6.2.3 Thematic Structure

The thematic structure of a text constitutes its thematic coherence in that it is responsible for the thematic connections between micro- and macrosegments of the text, and for their connection to an overall discourse topic, which serves as a frame for integrating the subtopics with regard to content. These connections between discourse topics and subtopics (and the thematically homogeneous macrosegments of the text, respectively) can be either semantic or functional or schema-based/associative. They constitute global thematic coherence.

Apart from thematic coherence on the global level, coherence can also be manifested by a relationship between adjacent sentences or clauses (i.e. elementary discourse segments). Such local relations are often signalled by explicit grammatical connections, which are formally realised by recurrence (e.g. coreference, anaphora) or by means of connectivity (e.g. conjunctions). These forms of connections are also called cohesion. Existing frameworks which model these local connections between elementary discourse segments operate on one of the different levels of discourse structure, i.e. referential structure (anaphoric relations), thematic structure (thematic development) and relational discourse structure (rhetorical relations).

The best known model for the description of local thematic development (i.e. the thematic relations between elementary discourse segments) is the *model of thematic progression* by Danes (1970). Another, similar, *model of thematic organisation* was proposed by Zifonun et al. (1997). Their proposed major patterns of local thematic development can be summarised as follows: 1. Continuation (of theme or rheme[3]) 2. Derived Theme (a. derived from a hypertheme, b. derived from a preceding theme or rheme), 3. Associated Theme. Apart from associated theme, all connections

between two adjacent topics are based on semantic relations like *part-of* or *identity* and are often explicitly signalled by means of coreference. But such connections are not sufficient to describe all possible thematic relations. As Brinker (1997) points out, models like the one by Danes (1970) do not cover anything that cannot be covered by an analysis of the referential structure alone.

Research investigating functional and associative connections between topics is therefore important to overcome limitations of models which solely focus on semantic or referential ties between sentences to describe patterns of thematic development. Examples of more functionally oriented research are Lötscher (1987), Brinker (1997) and Schröder (2003), who propose functional relations like *reason, justify*, or *exemplification* to model thematic connections. The integration of functional relations in the analysis of the thematic structure seems quite natural, because an elaboration of a topic not only comprises the elaboration of its parts (which could be modelled by semantic relations like hyperonymy) but also the specification of functionally connected aspects of the topic, which could be modelled by RST relations.

To be able to model both kinds of relations (semantic and functional) in one discourse representation framework, we interpret the RST relation ELABORATION to represent coherence relations between discourse segments that are induced by the semantically motivated relations between discourse referents contained in them. For a detailed modeling of patterns as described in Danes (1970) or Zifonun et al. (1997), an extension of the ELABORATION relation with different subtypes was necessary. Figure 6.3 shows the subtypes that we defined for discourse annotation in the project *SemDok*.[4]

ELABORATION-DERIVATION comprises all relations between a nucleus and a satellite which are based on topic derivation, or ontological subordination. The subtypes of this relation are all mentioned in various publications but have not been grouped together before (cf. Mann and Thompson 1988, Hovy and Maier 1995, Carlson and Marcu 2001). ELABORATION-IDENTITY holds between a nucleus and a satellite that share a referential identity, that are about the same discourse referent. On the one hand we distinguish between forms of *theme-theme-* or *rheme-theme-*chaining (cf. Polanyi et al. 2003), on the other hand between *assignment* (of a technical term or an abbreviation) and other forms of *specification* where the meaning of the topic in the nucleus is expanded, restricted or specified by a syntactically incomplete satellite.

With this extension of the set of rhetorical relations we can capture all patterns of thematic development by means of RST (Table 6.1). It must be emphasised that ELABORATION has some special characteristics compared with other discourse relations: First, it is a relation that potentially holds between all thematically connected discourse segments. It is therefore one of the "most prevalent forms of modification of a nucleus" and "extremely common at all levels of the discourse structure" (Carlson et al. 2001) – in our corpus, ELABORATION is the second most frequent relation (about 25% of all relation instances in the presently annotated subcorpus). In an annotation process ELABORATION can be overridden by more specific discourse relations, i.e. whenever there are signals for a more specific discourse

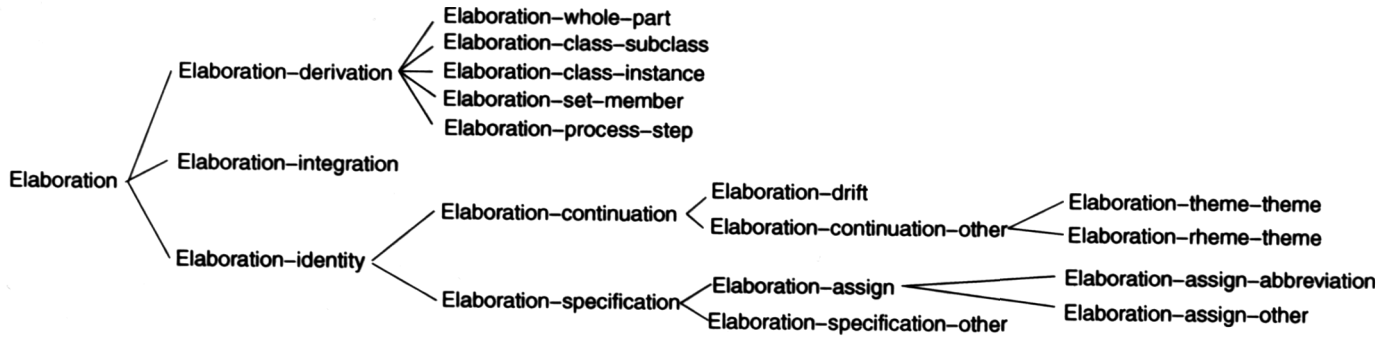

Fig. 6.3 SemDok hierarchy of ELABORATION relations

Table 6.1 Thematic relations

Patterns of thematic development	Thematic connections	
	Semantic relations	Rhetorical relations
(Referential) Continuation	synonymy, identity, paraphrase	ELABORATION-IDENTITY ELABORATION-CONTINUATION ELABORATION-SPECIFICATION ELABORATION-RESTATEMENT ELABORATION-EXAMPLE ELABORATION-DEFINITION
(Ontological) Derivation	hyponymy, hyperonymy, partonymy, meronymy	ELABORATION-DERIVATION ELABORATION-SET-MEMBER ELABORATION-PROCESS-STEP ELABORATION-CLASS-SUBCLASS ELABORATION-CLASS-INSTANCE ELABORATION-WHOLE-PART ELABORATION-INTEGRATION
(Functional) Supplementation/ Association		BACKGROUND, CIRCUMSTANCE CAUSE, RESULT, CONSEQUENCE PURPOSE, CONDITION, CONTRAST INTERPRETATION, EVALUATION,...

relation to hold between two discourse segments, this more specific relation is annotated. Second, ELABORATION is seldom signalled by syntactic or lexical discourse markers. Instead, ELABORATION may be identified by means of those linguistic features that signal thematic development: lexical-semantic and referential (anaphoric) relations between the central discourse entities of two discourse segments as well as lexical chains (Morris and Hirst 1991). As shown in Table 6.1, ELABORATION-DERIVATION and the converse relation ELABORATION-INTEGRATION are theoretically signalled by semantic relations like *hyponymy*, *hyperonymy*, *holonymy* etc., ELABORATION-IDENTITY by relations like *synonymy*, *identity* etc. Figure 6.4 and 6.5 show two examples where *holonymy* induces ELABORATION-DERIVATION, and *pertonymy* ELABORATION-DRIFT.[5]

These semantic relations (and the corresponding ELABORATION subtypes) can in principle be identified by consulting a lexico-semantic resource like *GermaNet* (cf. Kunze 2001) – only the coverage of GermaNet 5.0 is not sufficient for our corpus of scientific articles: only 69.3% of all noun tokens and 41.8% of all noun types in our corpus can be found in it (cf. Bärenfänger et al. 2007). We therefore primarily focus on the identification of ELABORATION and its subtypes by means of (annotations of) anaphoric relations and lexical chains as supplied by our project partners.

In various studies it has been pointed out that thematic development is closely connected with referential continuity, and that anaphoric relations may be used as signals for thematic continuity (cf. Danes 1970, Givon 1983, Zifonun et al. 1997). For the utilisation of anaphoric relations as cues for ELABORATION we cooperate with the *Sekimo* project where our corpus was annotated according to a schema for anaphoric relations (CHS, cf. Holler 2004). Two types of intra-textual anaphoric

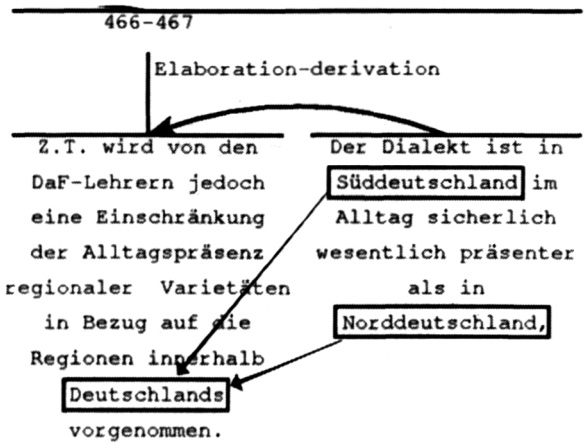

Fig. 6.4 Holonymy as a cue for ELABORATION-DERIVATION

relations are distinguished: *bridging* and *cospecification* relations. In cospecification relations (COSPEC), anaphora and antecedent are referentially identical, while bridging relations (BRIDGING) are based on semantic relations like meronymy, set-membership, and associative relations between anaphor and antecedent which have to be inferred from context.

Analyses of our corpus have shown that the presence of an anaphoric relation between discourse entities in two discourse segments is (approximately) a necessary condition for ELABORATION to hold between them. Yet, it is not a sufficient condition – this is amongst other things due to the status of ELABORATION as a default relation. However, correlations between certain subtypes of ELABORATION and specific anaphoric relations could be found as well, e.g. in 66.7% of all

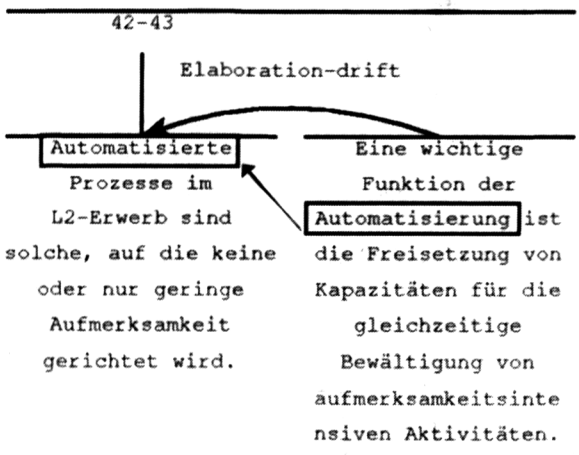

Fig. 6.5 Pertonymy as a cue for ELABORATION-DRIFT

```
<relationInstance rtype="elaboration-continuation-other">
    <segment1>Im folgenden Abschnitt werden wir zunächst einige terminologische Klärungen
            vornehmen .
    </segment1>
    <segment2>Diese betreffen einerseits unser Verständnis von regionalen Varietäten
            ( 2.1 ), andererseits das Spracheinstellungskonzept ( 2.2 ).
    </segment2>

    <anaphora atype="cospec:ident">
            <antecedent>einige terminologische Klärungen</antecedent>
            <anaphor>Diese</anaphor>
    </anaphora>
</relationInstance>
```

Listing 6.2 Correspondence of COSPEC:IDENTITY and ELABORATION-CONTINUATION

occurrences of *bridging:has-member*, ELABORATION-INTEGRATION holds, and in 82% of all ELABORATION-CONTINUATION occurrences, *cospec:ident* holds. An example of the latter is shown in Listing 6.2.[6]

Another approach to identifying thematically connected discourse segments is based on lexical cohesion, or, more specifically, the presence of lexical chains between discourse segments. "Lexical chains tend to indicate the topicality of segments" (Morris and Hirst 1991). This suggests that lexical chains can be employed to identify pairs of thematically homogeneous segments and, conversely, thematic breaks within logically defined segments. Lexical chains could thus also be used to revise the segment boundaries defined by the logical document structure. Incidents where discourse or thematic structure deviates from the logical document structure defined by the author of a text have sometimes been observed (cf. Stein 2003, Sporleder and Lapata 2004). In the two partner projects *HyTex* (see Storrer in this volume; Lenz in this volume) and *IndoGram* (Mehler in this volume), algorithms for the automatic construction of lexical chains have been implemented.

As emphasised above, thematic structure can be split into a local and a global level. Using RST, it is possible to analyse and represent both levels, the local level by annotating the relations between adjacent elementary discourse segments and the global level by relating complex discourse segments. Particularly for the analysis of the latter relations across larger spans of text, the relation ELABORATION and its subtypes are beneficial (cf. also Carlson et al. 2001). The goal of our approach to thematic structure is thus not to identify and label discourse topics, but to integrate semantic and functional thematic relations in one discourse representation model.

6.2.4 Generic Document Structure

Genre-specific *superstructure* or *text type structure* (van Dijk 1980, Swales 1990) is an aspect of global discourse structure. An analysis of our corpus showed that most scientific articles are sequentially structured along the text type-specific categories *problem, evidence, answers*, although deviations are possible, and commonly found (cf. Bärenfänger et al. 2006). These text type-specific functional categories (also e.g. *method, results,* and *discussion*) can be hierarchically organised in a text type

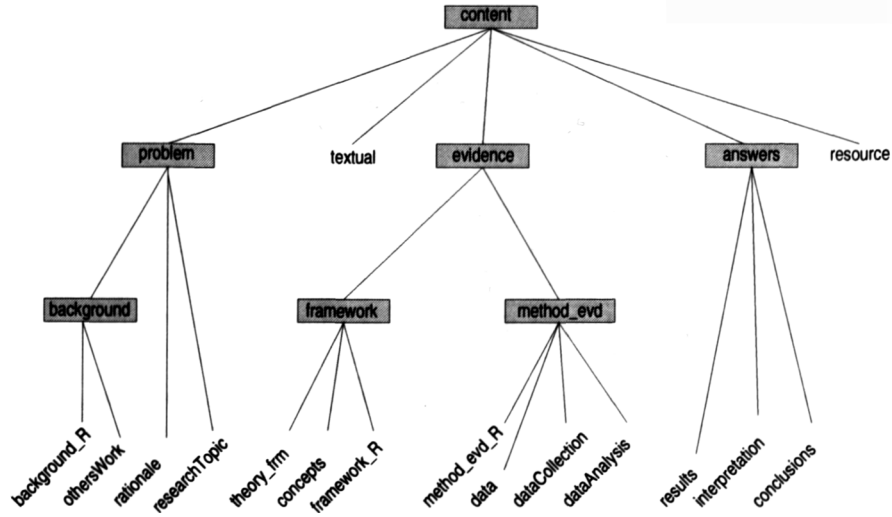

Fig. 6.6 Text type structure (TTS) schema (23 categories)

structure schema. One such schema (cf. Fig. 6.6) was designed in the first phase of the present project and is used for the *text type structure* corpus annotation level (TTS) described in Section 6.3.2. Previous approaches to text parsing of scientific articles have focussed on automatically assigning text type-specific functional categories (or *zones*, after Teufel 1999) from the text type structure to text segments using automatic text categorisation methods (Kando 1999, Teufel and Moens 2002, Langer et al. 2004a).

One aim of the present project, however, is to formulate a method to integrate text type structure and overall relational discourse structure. Text structural categories are functions of text parts within the whole text, i.e. they represent a mapping between pairs of one text span and the whole text into the set of textual category labels. RST analyses can be viewed as functions that map pairs of text spans onto a rhetorical relation label. Several of the category names used in previously proposed text type schemas (Kando 1999, Teufel and Moens 2002, Langer et al. 2004a) such as *problem, results, conclusion* suggest that text type structure and rhetorical structure can actually be interleaved (cf. Gruber and Muntigl 2005). This hypothesis

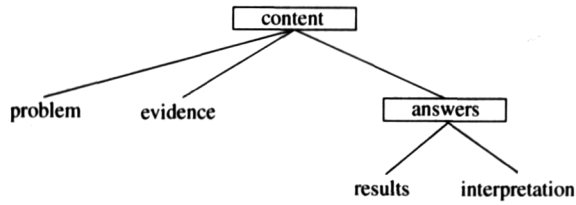

Fig. 6.7 Possible instantiation of text structural categories

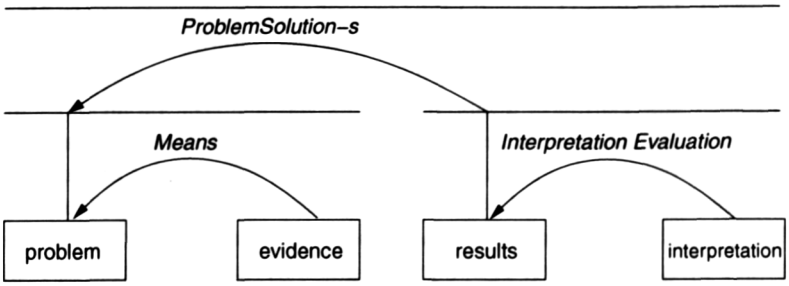

Fig. 6.8 Relational structuring of the categories in Fig. 6.7

is supported by the results of an empirical analysis of our corpus which showed significant correlations between generic and rhetorical structure. An *interpretation* constituent in a text type structure schema instantiation of an article (Fig. 6.7) can, for example, very often be characterised as an RST satellite to a nucleus which are related through INTERPRETATION (Fig. 6.8). The distribution of RST relations over the different TTS categories shows clear deviations from a normal distribution – some TTS and RST pairings are much more likely to occur than other pairings, e.g. the TTS category *OthersWork* significantly correlates with the RST relation BACKGROUND, *ResearchTopic* with ELABORATION. The overall findings of the corpus study are described in full length in Bärenfänger et al. (2006).

6.3 Resources

6.3.1 Corpus

For the development of the knowledge sources and the preprocessing components of the discourse parser, we work with a corpus that was compiled and annotated during the first project phase (2001–2004). The corpus comprises 120 scientific articles from two different disciplines (psychology and linguistics), languages (English and German) and sub-genres (experimental and review). English psychological and linguistic documents were taken from electronically available journals which were ranked highly in the listings of the Institute for Scientific Information (ISI) and published in the years 2000–2002. German linguistic articles were compiled from the online-journal "Linguistik Online" (volumes 2000–2003).

6.3.2 Annotation Levels

Our approach to corpus annotation was based on the assumption of four annotation levels that play a role in discourse analysis. (a) logical document structure (as e.g. encoded in DocBook, cf. Walsh and Muellner 1999, or *ldoc*, cf. Stede and

Suriyawongkul in this volume), (b) genre-specific text type structure (as described in van Dijk 1980, Swales 1990, Kando 1999, Teufel and Moens 2002), (c) rhetorical structure (Mann and Thompson 1988), and (d) syntactic structure. To examine dependencies between these levels, the corpus was analysed on all of them, and the analyses themselves were represented as XML-based multi-layer-annotations (Witt et al. 2005). In the multi-layer annotation approach, each information level is realised as an independent XML annotation layer and stored in a separate file. Thus, we distinguish between annotation *levels* (abstract information levels such as the syntax and morphology level of a linguistic grammar) and annotation *layers* (their realisations in XML) (cf. Goecke et al. in this volume). In the following, the levels and XML layers of logical document structure, text type structure, and rhetorical structure are described in more detail.

Logical document structure (DOC): The logical document structure is an abstraction of the physical layout structure. The annotation of the logical document structure (abbreviated DOC) – i.e. the hierarchical division of the text in sections, titles, paragraphs, footnotes, lists etc. – was provided using a subset of the DocBook DTD, extended by 13 elements relevant for the corpus (such as `<footnoteSect>`).

Text type structure (TTS): To represent the canonical text type structure of a scientific article (see Section 6.2.4), an XML schema was created which contains 135 functional categories such as *framework*, *method*, or *dataCollection*. The creation of the text type schema was based on an empirical analysis of the corpus and on an evaluation of similar approaches regarding so-called rhetorical zones (Teufel and Moens 2002) and text-level constituents (Kando 1999). The categories are arranged hierarchically in the schema. The resulting tree structure was also used to generate a reduced schema with 23 categories, which is more suitable for an efficient and consistent annotation. Besides, as linguistic articles show a variety of orders of functional categories, a flat schema version was derived from the hierarchical one by means of an XSLT style sheet. Articles annotated according to the flat schema still contain information about the original hierarchical structure encoded using the ID/IDREF-mechanism of XML (cf. Bayerl et al. 2003a, Langer et al. 2004a).

Rhetorical structure (RST): The rhetorical structure describes functional-argumentative relations (e.g. CONCESSION, or EVIDENCE) between discourse segments, cf. Section 6.2.2. The set of rhetorical relations used for the annotation of the corpus is basically the one proposed by Mann and Thompson (1988) in the framework of Rhetorical Structure Theory (RST). We employed the RSTTool developed by O'Donnell (2000) to manually annotate the rhetorical structure. By means of a Perl program, we can convert the flat XML output of the RSTTool to our hierarchical *RST-HP-format*, which, together with some extensions will be the format of the target structure of our discourse parser, cf. Lüngen et al. (2006a). From the English psychological articles, 15 sections (2–3 pages each) were annotated starting from

elementary discourse segments, and 10 German linguistic articles were annotated completely but starting from paragraphs as smallest units. Currently, the rhetorical annotations are being extended using the more scenario-specific relation set RRSET described in Section 6.3.3.2. The RST annotations serve as training and evaluation material for the discourse parser.

Syntactic structure (CNX): The morphology/syntax layer was created automatically using the commercial *Machinese Syntax* tagger software from Connexor Oy.

During the annotation process, the quality of the manual annotations was supervised in two ways: Inter-rater reliability and intra-individual consistency (coder drift) were checked for the manually created annotations (cf. Bayerl et al. 2003b) using κ as a measure of agreement (Cohen 1960). The results of the tests for inter-rater reliability show that the quality of the TTS annotation was "substantial" (average $\kappa = .64$). κ for the RST annotations was .77 for the intra-sentential relations. The quality of the DOC annotation ($\kappa = .98$) is "nearly perfect" (cf. Landis and Koch 1977).

Table 6.2 Corpus annotations

	TTS (135)	TTS (23)	DOC	RST	CNX
English psychological articles	73	73 (automatically generated)	73	15 (several sections)	73
German linguistic articles		47	47	3 + 10 CDS-block	47

The extensive XML-based multi-layer-annotated corpus gives us the possibility to examine interrelations between these levels and to identify cues for rhetorical relations, e.g. cues on the level of document structure (such as an occurrence of the element `<itemizedList>`) or syntactic or topical cues (e.g. the occurrence of the text type-category *dataCollection*). Moreover, cues from different annotation levels can be combined to form complex conditions for the assignment of a specific rhetorical relation.

6.3.3 Additional Resources

6.3.3.1 Discourse Marker Lexicon

Discourse markers are functional elements that can be regarded as signals for a rhetorical relation (coherence relation) between two text segments. As we have indicated above, there are different types of discourse markers: Firstly, there are lexical discourse markers, or *connectives*. These are syntactically mostly adverbs or conjunctions. They may consist of one word (*weil*, "because"), multiple adjacent

parts (*so dass*, "so that") or multiple discontinuous parts (*wenn . . . dann . . . sonst . . .*, "if . . . then . . . else . . ."). Secondly, configurations of grammatical and/or document type-related features can function as (more abstract) discourse markers. An occurrence of a `<doc:itemizedlist>`-environment on the logical document structure level would indicate one nucleus of a multinuclear LIST or SEQUENCE relation, `<doc:glossterm>` would induce the nucleus of an ELABORATION-DEFINITION relation, `<doc:glossdef>` its satellite, and `<doc:title>` the satellite of a PREPARATION relation. In the present stage of the project, the lexicon comprises lexical discourse markers, other discourse markers are currently treated in the rule component of the parser.

Many lexical discourse connectives are highly ambiguous. Frequently they do not clearly denote an individual rhetorical relation, but on the contrary the same markers signal different relations depending on their context. Our intention was to provide an XML-encoded inventory of German discourse connectives which resolves these ambiguities.

First, we extracted a list of discourse connectives from our corpus and developed a suitable representational format in XML. The definition and validation of the XML data was implemented in XML-Schema. The dictionary contains orthographic and syntactic characteristics of the respective discourse markers. The syntactic information included is based on the annotation generated by the *Machinese Syntax* Tagger from *Connexor Oy*, the descriptions in the *Handbuch der deutschen Konnektoren* of the *IDS Mannheim* (Pasch et al. 2003) and the grammar by Helbig and Buscha (1998). The encoding the topological fields resembles the format employed in DiMLex (Stede and Umbach 1998).

```
<dm id="c63" typ="lexical">
  <cue>
    <text>wenn</text>
    <lemma pos="CS">wenn</lemma>
    <position>
      <sub>+</sub>
    </position>
  </cue>
  <kommentar>"wenn auch X" is always Concession.</kommentar>
  <kommentar>"wenn X auch" is an alternative - not yet considered here</kommentar>
  <filter>
    <!-- obligatory conditions -->
    <hypothese relname="Concession">
      <word fenster="9" richtung="r">
        <text>auch</text>
        <lemma pos="ADV">auch</lemma>
      </word>
    </hypothese>
  </filter>
  <rels default="Circumstance">
    <relation score="0.5" relname="Circumstance" skopus="eds+" typ="s" beds-richtung="lr"/>
    <relation score="0.5" relname="Concession"   skopus="eds+" typ="s" beds-richtung="lr">
      <!-- optional conditions -->
    </relation>
  </rels>
</dm>
```

Listing 6.3 Entry for "wenn" in the discourse marker lexicon

Each entry in the dictionary is represented by a <dm>-element (see Listing 6.3 for a sample entry). A <dm>-entry generally consists of three main parts: an identification unit, a filter unit, and an allocation unit. The identification unit identifies a lexical discourse marker (word or phrase) by its form (<text>), by the word stem (<lemma>) and its part of speech (@pos). The optional filter unit allows for disambiguation of discourse markers by providing hypotheses about possible contexts and their associated specific rhetorical relations. Obligatory combinations of features (of the current segment and the reference segment) are combined to form hypotheses. Its attributes are supposed to override the general attribute values given in the allocation unit with their specific values in the current context. In the allocation unit all relations expressed by the discourse marker are specified. The @score attribute contains the conditional score for the relation given the discourse marker. It is presently based on an assumption of equal distribution but will eventually be estimated from our corpus. The attributes @beds-richtung and @skopus determine the position of the segment in comparison to the reference segment and the scope of the segment. If a segment offers several competing relations signalled by different discourse markers, a hierarchy of relations can be expressed on the basis of the attribute @skopus, so that the discourse parsing engine has criteria for a decision about the order in which the individual relations are applied and promoted (cf. Corston-Oliver 1998). The allocation unit can contain additional conditions for the segment and the reference segment, which provide the discourse parsing engine with further indicators to confirm a relation or to find the corresponding reference segment.

The discourse marker lexicon currently contains 92 <dm> entries. A perl program for tagging lexical discourse markers in texts based on a CNX annotation of the text (see Section 6.3.2) and the discourse marker lexicon exists.

6.3.3.2 Set of Rhetorical Relations

One goal of the present project was to develop a set of rhetorical relations suitable for analysing scientific articles in our explorative reading scenario, cf. Section 6.1. Our strategy was as follows: We took the extended classical MT (Mann/Thompson) relation set of 34 relation types as a starting point (cf. Mann and Taboada 2005); additionally we reviewed the comprehensive relation taxonomies previously suggested for English by Carlson et al. (2001) (96 relation types, 78 of which are at the base level of the taxonomy, which were employed in the rhetorical analysis of newspaper articles) and Hovy and Maier (1995) (65 relation types, 43 of which at base level, which were designed mostly from the perspective of natural language generation and are not RST-specific) and chose candidate relations for extending the MT relation set. We then evaluated the RST annotations that were available from the first project phrase (see Section 6.3.1) for determining the relevance of each relation in our corpus. Subsequently, we designed our relation set (called the RRSET) along the following criteria:

- we introduced subrelations when we found strong associations with certain discourse markers that seemed highly scenario-relevant; for instance we wanted to

distinguish between LIST-COORDINATION relations that come about by syntactic
coordination vs. LIST-DM_OTHER relations that come about through discourse markers on the logical document structure level such as the <listitem> elements. Similarly, we introduced PREPARATION-TITLE, PREPARATION-QUESTION, Preparation-other, CITATION-EVIDENCE, and CITATION-ATTRIBUTION;

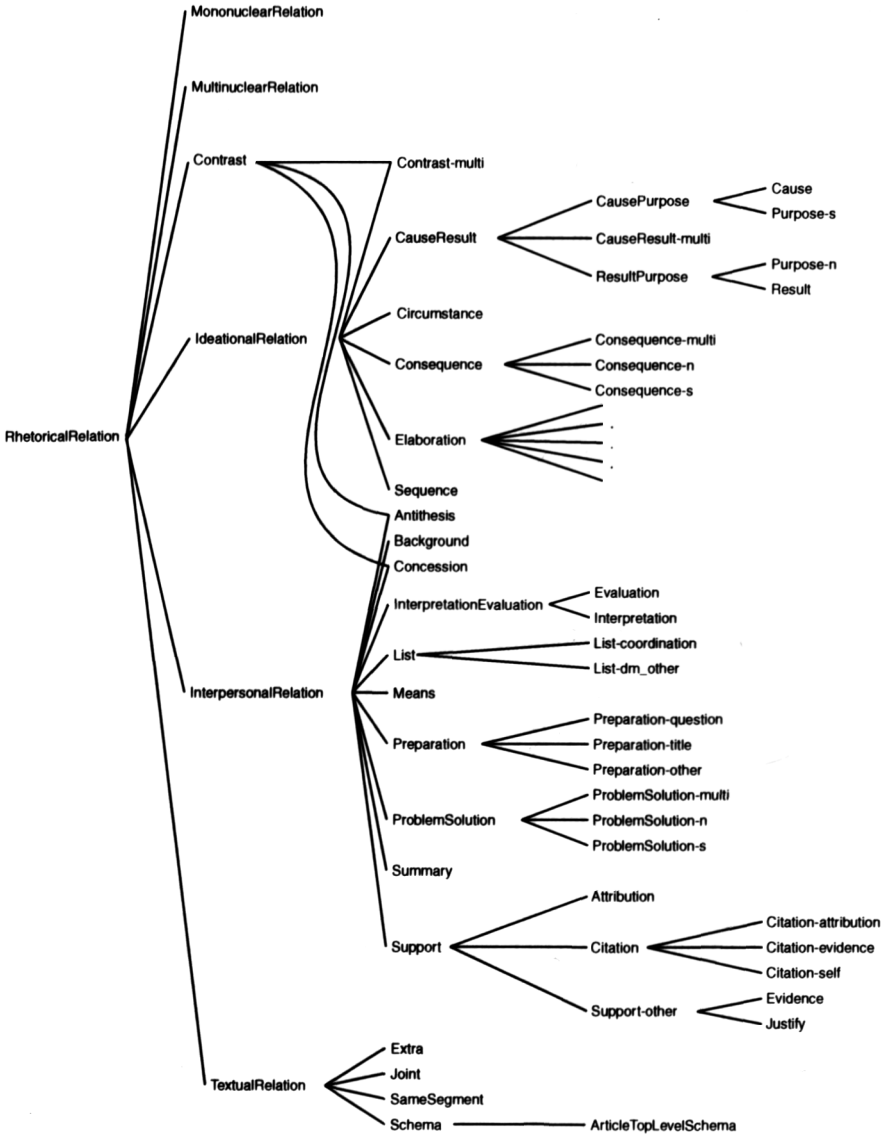

Fig. 6.9 SemDok RRSET ontology (save the subclasses of ELABORATION) (edges from MONONUCLEARRELATION and MULTINUCLEARRELATION are not shown)

- we introduced the comprehensive sub-taxonomy of ELABORATION relations described in Section 6.2.3;
- we omitted two relations from Mann and Thompson (1988), which had proved to be irrelevant in our text type (MOTIVATION and ENABLEMENT);
- we introduced new superordinate relation classes for relations that were hardly distinguished by discourse markers and that were also often confused by human annotators when trying to apply semantically oriented definitions (SUPPORT-OTHER, CONTRAST, LISTSEQUENCE, and INTERPRETATIONEVALUATION);
- we introduced relation types that denote heavily underspecified relations (MONONUCLEARRELATION, MULTINUCLEARRELATION, IDEATIONALRELATION, INTERPERSONALRELATION, and TEXTUALRELATION).
- we introduced certain subrelations based on alternative nuclearity assignments as in Carlson et al. (2001) (CONSEQUENCE-MULTI, CAUSERESULT-MULTI, and PROBLEMSOLUTION-MULTI);

The resulting *SemDok* RRSET taxonomy consists of 70 relation types (44 at base level) and is encoded in the semantic web ontology language OWL (cf. Bechhofer et al. 2004 and see also Farrar and Langendoen in this volume). OWL consists of the three sublanguages OWL Lite, OWL DL and OWL Full, which differ in expressivity. We chose OWL DL (based on *d*escription *l*ogics) to encode our RRSET ontology because current reasoning software such as RacerPro,[7] which can be used for consistency checking and drawing inferences, is designed for the decidable sublanguage OWL DL. Since we wanted to declare disjointness between certain rhetorical relation types and to encode properties of rhetorical relations that are to be inherited by their subrelations, we modelled RST relations as OWL classes. All RRSET relations are cross-classified along the two dimensions *nuclearity* and *metafunction*, giving rise to multiple inheritance. SUPPORT, for instance, is both a subclass of INTERPERSONALRELATION and of MONONUCLEARRELATION. Figure 6.9 shows the hierarchy of relations as induced by the `<rdfs:subClassOf>` specifications of the OWL representation (the complete ELABORATION subhierarchy is shown in Fig. 6.3) (cf. Bärenfänger et al. 2007).

6.4 Discourse Parsing Architecture

This section shortly describes the architecture and algorithm of the discourse parser that is developed in the *SemDok* project based on the theoretical assumptions and resources described so far. In Fig. 6.10, the architecture of the discourse parsing system is shown.

The declarative knowledge sources are used in several preprocessing steps by auxiliary components to analyse an input text[8] and to provide it with multiple annotation layers. The discourse parser itself takes these different annotation layers as its input to guide its decisions for selecting and applying rules to build up a set of possible annotations of the rhetorical discourse structure.

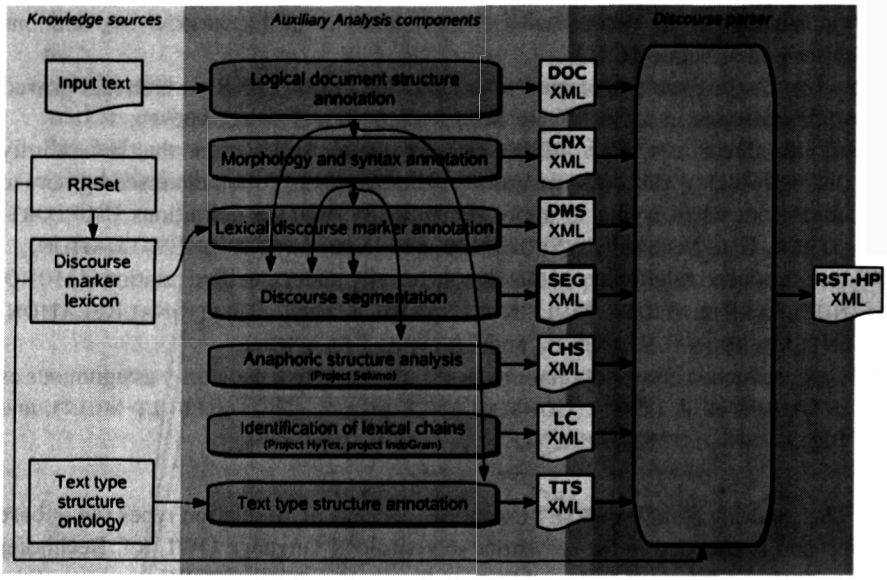

Fig. 6.10 Discourse parser architecture

The base annotation layer controlling the parsing cycles is the discourse segmentation layer (SEG, cf. Section 6.2.2). The parser strategy follows the segmentation bottom-up, firstly combining adjacent elementary discourse segments (EDS) recursively to sentential ones (SDS), secondly combining sentential discourse segments to complex discourse segments (CDS) of type block, then combining the block level segments to division level segments (i. e. sections) and finally division level segments up to the level of the complete document. Each one of these phases is called a *cascade step*.

The remaining annotation layers – i. e. the logical document structure (DOC), the morphological and syntactic tagging (CNX, cf. Section 6.3.2), the lexical discourse marker annotation (DMS, cf. Section 6.3.3.1) and the anaphoric structure (CHS, cf. Section 6.2.3) – provide linguistic cues and constraints for rhetorical relations and are referenced in the rule component of the parser. These cues and constraints describe correlations between different discourse markers represented as configurations of XML elements and attributes on the different annotation layers and yield hypotheses of rhetorical relations that hold between the discourse segments of the input text.

The output of the parser is a set of well-formed RST trees in the extended RST-HP format described in Lüngen et al. (2006a).

The parsing strategy used in each cascade step is a bottom-up passive chart parser. Each time a tuple of adjacent spans matches all cues and constraints specified in a rule, a new edge is inserted into the chart, labelled with a rhetorical relation and a nuclearity setting, and representing a new discourse-coherent span over the input segments.

To evaluate competing rhetorical relation hypotheses, each edge of the chart is assigned a score expressing its adequacy in a resulting discourse structure. The score of an edge depends on the context in which it is inserted into the chart, i. e. it is a function of the scores of its child edges and the score of the rule which is applied to insert the edge (cf. Le Thanh et al. 2004). The rule score is composed of the a priori probability of the rhetorical relation that is induced by the rule, i. e. the probability with which the relation occurs in the corpus, combined with the conditional probability of the relation given the discourse marker that is mentioned in the rule. The probabilities are estimated by calculating percentages of occurrences of relation instances and discourse markers in the development corpus.

To reduce the search space, a list of *applied discourse marker identifiers* is also associated with each edge in the chart. They form a control structure that ensures that one discourse marker cannot be applied twice during the construction of a rhetorical tree. Thus, they partly replace the promotion sets as proposed by Marcu (2000).

In forthcoming versions of the parser, the processing of additional linguistic resources shall be incorporated. These are annotations of lexical chains (cf. Section 6.2.3) and genre-specific text type structure (cf. Section 6.2.4).

6.5 Conclusion

In this chapter, we discussed the theoretical foundations of discourse analysis of texts of a complex genre from the perspectives of document engineering, discourse theory, and text linguistics. We identified those aspects of discourse and discourse analysis that are relevant for our text type and application scenario, especially the prominent role of logical document structure, thematic structure and text type structure when analysing a complex genre. We argued that for discourse analysis, an augmented version of RST (Rhetorical Structure Theory) should be adopted. One of the proposed augmentations is the accomodation of thematic structure by differentiating various subtypes of the familiar ELABORATION relation based on semantic relations between the themes of the discourse.

Subsequently, the resources and methods based on XML technologies that we use in developing the discourse parser were introduced in more detail. We sketched the structure of the discourse marker lexicon which contains mostly lexical discourse markers (connectives). We introduced a set of 44 rhetorical relation labels based on the original RST relation set but adapted to our project scenario and text type. The rhetorical relation labels are hierarchically ordered in a relation taxonomy.

One future focus of the project will be on evaluation. The output of the system will be compared to manually created annotations of our corpus, which will serve as a gold standard, using standard methods and measures. Besides, system performance will be compared with a baseline provided by trivial algorithms such as random relation assignment or exclusive assignment of the most frequent relation

according to our corpus annotations. Moreover, it will be interesting to compare the system with other existing RST parsers, such as the one for German newspaper commentaries developed at Potsdam University (cf. Reitter 2003a, b).

Traditionally, the external evaluation of a discourse parser is done by assessing its contribution to automatic text summarisation systems (cf. Marcu 2000, Rehm 1998, Polanyi et al. 2004a). In the context of the DFG-Forschergruppe *Texttechnologische Informationsmodellierung*, however, it will be possible to examine whether and how the analyses provided by the discourse parser can improve the performance of the automatic hypertextualisation system developed in the *HyTex* project (Lenz in this volume, Storrer in this volume).

Notes

1. Text extract from Saari, Mirja (2000). Schwedisch als die zweite Nationalsprache Finnlands: Soziolinguistische Aspekte. *Linguistik Online*, 7, http://www.linguistik-online.de.
2. We employ the tool described in O'Donnell (2000) for drawing RST trees.
3. *Rheme* as in the theory of *Functional Sentence Perspective.*
4. Apart from the subtypes of elaboration shown in Fig. 6.3, we distinguish ELABORATION-EXAMPLE, ELABORATION-DEFINITION and ELABORATION-RESTATEMENT.
5. Text extract from Baßler, Harald and Helmut Spiekermann (2001). Dialekt und Standardsprache im DaF-Unterricht. Wie Schüler urteilen – wie Lehrer urteilen. *Linguistik Online* 9, http://www.linguistik-online.de.
6. Extract from Baßler and Spiekermann (2001) (see Footnote 5).
7. http://www.racer-systems.com
8. The test corpus consists of a suite of German linguistic journal articles.

References

Asher, Nichoas and Vieu, Laure (2005). Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Asher, Nicholas and Lascarides, Alex (2003). *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Bärenfänger, Maja, Lüngen, Harald, Hilbert, Mirco, and Lobin, Henning (in press). The role of logical and generic document structure in relational discourse analysis. In Benz, Anton, Kühnlein, Peter, and Sidner, Candy, editors, *Constraints in Discourse 2*. Series Pragmatics & Beyond. John Benjamins, Amsterdam.

Bärenfänger, Maja, Lobin, Henning, Lüngen, Harald, and Hilbert, Mirco (2008). OWL ontologies in discourse parsing. *LDV-Forum. GLDV-Journal for Computational Linguistics and language Technololgy* 23(1):7–26.

Bayerl, Petra Saskia, Lüngen, H., Gut, U., and Paul, K.I. (2003a). Methodology for reliable schema development and evaluation of manual annotations. In *Workshop Notes for the Workshop on Knowledge Markup and Semantic Annotation, Second International Conference on Knowledge Capture (K-CAP 2003)*, pages 17–23, Sanibel, Florida.

Bayerl, Petra Saskia, Lüngen, Harald, Goecke, Daniela, Witt, Andreas, and Naber, Daniel (2003b). Methods for the semantic analysis of document markup. In *Proceedings of the ACM Symposium on Document Engineering (DocEng 2003)*, pages 161–170, Grenoble.

Bechhofer, Sean, van Harmelen, Frank, Hendler, Jim, Horrocks, Ian, McGuiness, Deborah L., Patel-Schneider, Peter F., and Stein, Andrea Lynn (2004). OWL Web Ontology Language – Reference. Technical report, W3C (World Wide Web) Consortium. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

Brinker, Klaus (1997). *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. 4th edition, Erich Schmidt, Berlin.

Carlson, Lynn and Marcu, Daniel (2001). Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.

Carlson, Lynn, Marcu, Daniel, and Okurowski, Mary Ellen (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Denmark.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46.

Corston-Oliver, Simon (1998). *Computing of Representations of the Structure of Written Discourse*. PhD thesis, University of California, Santa Barbara.

Daneš, Frantisek (1970). Zur linguistischen Analyse der Textstruktur. *Folia Linguistica*, 4:72–78.

Danlos, Laurence (2005). Comparing RST and SDRT discourse structures through dependency graphs. In Sassen, Claudia, Benz, Anton, and Kühnlein, Peter, editors, *Proceedings of Constraints in Discourse*, pages 55–62, Dortmund.

Egg, Markus and Redeker, Gisela (2005). Underspecified discourse representation. In Sassen, Claudia, Benz, Anton, and Kühnlein, Peter, editors, *Proceedings of Constraints in Discourse*, pages 46–53, Dortmund.

Givon, Talmy (1983). Topic Continuity in Discourse: An Introduction. In Givon, Talmy, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, pages 5–41. John Benjamins, Amsterdam, Philadelphia.

Goecke, Daniela, Lüngen, Harald, Sasaki, Felix, Witt, Andreas, and Farrar, Scott (2005). GOLD and discourse: Domain- and community-specific extensions. In *Proceedings of the 2005 E-MELD-Workshop*, Boston, MA.

Gruber, H. and Muntigl, P. (2005). Generic and rhetorical structures of texts: Two sides of the same coin? *Folia Linguistica. Special Issue: Approaches to Genre*, XXXIX(1–2):75–114.

Helbig, Gerhard and Buscha, Joachim (1998). *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. 18th edition, Langenscheidt, Leipzig.

Holler, Anke und Jan Frederik Maas und Angelika Storrer (2004). Exploiting coreference annotations for text-to-hypertext conversion. In *Proceedings of LREC*, volume II, pages 651–654, Lisboa.

Hovy, Eduard and Maier, Elisabeth (1995). Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper, http://www.isi.edu/natural-language/people/hovy/publications.html.

Kando, Noriko (1999). Text structure analysis as a tool to make retrieved documents usable. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, pages 126–135, Taipei, Taiwan.

Kunze, Claudia (2001). Lexikalisch-semantische Wortnetze. In Carstensen, Kai-Uwe et al., editor, *Computerlinguistik und Sprachtechnologie: eine Einführung*, pages 386–393. Spektrum Verlag, Heidelberg.

Landis, J.R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Langer, Hagen, Lüngen, Harald, and Bayerl, Petra Saskia (2004a). Text type structure and logical document structure. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pages 49–56, Barcelona.

Le Thanh, Huong, Abeysinghe, Geetha, and Huyck, Christian (2004). Generating discourse structures for written texts. In *Proceedings of COLING'04*, Geneva, Switzerland.

Lötscher, Andreas (1987). *Text und Thema. Studien zur thematischen Konstituenz von Texten*. Reihe Germanistische Linguistik, 81. Niemeyer, Tübingen.

Lüngen, Harald, Lobin, Henning, Bärenfänger, Maja, Hilbert, Mirco, and Puskàs, Csilla (2006a). Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB)*, pages 247–256, Bansko, Bulgaria.

Lüngen, Harald, Puskàs, Csilla, Bärenfänger, Maja, Hilbert, Mirco, and Lobin, Henning (2006b). Discourse segmentation of German written text. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, pages 245–256, Åbo, Finland. Springer.

Mann, William C. and Taboada, Maite (2005). RST – Rhetorical Structure Theory. W3C page. http://www.sfu.ca/rst.

Mann, William C. and Thompson, Sandra A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281.

Marcu, Daniel (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th annual meeting of the ACL*, pages 365–372, Maryland. Association for Computational Linguistics.

Marcu, Daniel (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

Morris, Jane and Hirst, Graeme (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Motsch, Wolfgang and Viehweger, Dieter (1991). Illokutionsstruktur als Komponente einer modularen Textanalyse. In Brinker, Klaus, editor, *Aspekte der Textlinguistik*, volume 106/107 of *Germanistische Linguistik*, pages 107–132. Olms, Hildesheim/Zürich/New York.

O'Donnell, Michael (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel.

Pasch, Renate, Brauße, Ursula, Breindl, Eva, and Waßner, Ulrich Hermann, editors (2003). *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Schriften des Instituts für Deutsche Sprache. de Gruyter, Berlin.

Polanyi, Livia, Culy, Chris, van den Berg, Martin, Thione, Gian Lorenzo, and Ahn, David (2004a). A rule based approach to discourse parsing. In *Proceedings of the 5th Workshop in Discourse and Dialogue*, pages 108–117, Cambridge, MA. 2004.

Polanyi, Livia, Culy, Chris, van den Berg, Martin, Thione, Gian Lorenzo, and Ahn, David (2004b). Sentential structure and discourse parsing. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pages 49–56, Barcelona.

Polanyi, Livia, van den Berg, Martin, and Ahn, David (2003). Discourse structure and sentential information structure. *Journal of Logic, Language and Information*, 12:337–350.

Rehm, Georg (1998). Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML- und DSSSL-basierten Repräsentation von RST-Relationen. Master's thesis, Universität Osnabrück.

Reitter, David (2003a). Rhetorical analysis with rich-feature support vector models. Master's thesis, University of Potsdam.

Reitter, David (2003b). Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. In Seewald-Heeg, Uta, editor, *Sprachtechnologie für die multilinguale Kommunikation. Textproduktion, Recherche, Übersetzung, Lokalisierung. Beiträge der GLDV-Frühjahrstagung 2003*, volume 18 of *LDV-Forum*, pages 38–52, Köthen.

Schröder, Thomas (2003). *Die Handlungsstruktur von Texten. Ein integrativer Beitrag zur Texttheorie*. Gunter Narr, Tübingen.

Sporleder, Caroline and Lapata, Mirella (2004). Automatic paragraph identification: A study across languages and domains. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 72–79, Barcelona.

Stede, Manfred and Umbach, Carla (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th international conference on Computational Linguistics (COLING-98)*, pages 1238–1242, Montreal, Canada.

Stein, Stephan (2003). *Textgliederung. Einheitenbildung im geschriebenen und gesprochenen Deutsch: Theorie und Empirie*, volume 69 of *Studia Linguistica Germanica*. de Gruyter, Berlin.

Swales, John M. (1990). *Genre Analysis. English in academic and research settings*. Cambridge University Press, Cambridge, UK.

Teufel, Simone (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh.

Teufel, Simone and Moens, Marc (2002). Summarizing scientfic articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

van Dijk, Teun A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Walsh, Norman and Muellner, Leonard (1999). *DocBook: The Definitive Guide*. O'Reilly, Sebastopol, CA.

Witt, Andreas, Lüngen, Harald, Goecke, Daniela, and Sasaki, Felix (2005). Unification of XML documents with concurrent markup. *Literary and Linguistic Computing*, 20(1):103–116.

Wolf, Florian and Gibson, Edward (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288.

Zifonun, Gisela, Hoffmann, Ludger, and Strecker, Bruno (1997). *Grammatik der deutschen Sprache*, volume 7 of *Schriften des Instituts für deutsche Sprache*, chapter C6 "Thematische Organisation von Text und Diskurs", pages 535–591. de Gruyter, Berlin/New York.